# The significance of scope in modelling tones in Chinese

*Branislav Gerazov*[1,2]*, Gérard Bailly*[2] *and Yi Xu*[3]

[1] FEEIT, University of Ss. Cyril and Methodius in Skopje, Macedonia
[2] Univ. Grenoble-Alpes, CNRS, Grenoble-INP, GIPSA-lab, 38000 Grenoble, France
[3] Department of Speech, Hearing and Phonetic Sciences, University College London, UK

gerazov@feit.ukim.edu.mk, gerard.bailly@gipsa-lab.fr, yi.xu@ucl.ac.uk

## Abstract

The Superposition of Functional Contours (SFC) prosody model decomposes the intonation and duration contours into elementary contours that encode specific linguistic functions. It can be used to extract these functional contours at multiple linguistic levels. The PySFC system, which incorporates the SFC, can thus be used to analyse the significance of including the neighbouring syllables in the scope of the tone functional contours in spoken Chinese on the modelling of prosody. Our results show that significant improvements of modelling tone functional contours are obtained by including the right syllable in the scope, but not the left one. We thus show that there is a larger carry-over effect for Chinese tones in contrast to an anticipatory one. This finding is in line with the established state-of-the-art.

**Index Terms**: tones, Chinese, scope, pitch, intonation, SFC

## 1. Introduction

Mandarin is a tone language in which every syllable carries a tone that can distinguish words that are segmentally identical [1, 2]. To model Mandarin intonation, it is critical to generate fine details of F0 contours associated with the lexical tones.

The Superposition of Functional Contours (SFC) model is a prosodic model that can decompose the prosody, including the pitch contour, into functionally relevant elementary contours [3, 4, 5]. It has been successfully used to model different linguistic levels, including: attitudes [6], grammatical dependencies [7], cliticisation [5], focus [8], as well as tones in Mandarin [9]. The experiments were carried out using PySFC, which is a prosody analysis system based on the SFC model, implemented in the scientific Python ecosystem [10]. The system has been licensed as free software and is available on GitHub[1].

In this paper we use the PySFC system to analyse the significance of the scope in modelling the functional contours associated with the four Mandarin tones. First we define the meaning of scope within the SFC modelling framework. In following, we discuss the state-of-the-art, our objectives and hypothesis. Finally we present our experimental framework the results, and a discussion thereof.

## 2. Scope in the SFC model

The functional contours in the SFC model are realised using neural network contour generators (NNCGs) [11, 6] that are trained in an analysis-by-synthesis loop that distributes the error at every iteration among the contour generators [12, 4]. Each NNCG is responsible for encoding a single linguistic function at different scopes, i.e. the collection of rhythmical units (RUs), e.g. syllables, it spans.

[1] https://github.com/bgerazov/PySFC

For each function that appears in a given utterance, the corresponding NNCG is used to generate the elementary contours for the different scopes that the function appears in. When all of the elementary contours in the utterance are overlapped and summed, they yield an approximate reconstruction of the multiparametric prosody contour of the original utterance. An example of this can be seen in Fig. 2, where the intonation contour of a Chinese utterance is decomposed into constituent elementary contours. The pitch here is expressed in 10s of quartertones. We can see that the SFC is powerful enough to correctly extract the pitch patterns of the four Chinese tones. It also captures the phrase final pitch drop (DC), and pitch reset in independent clauses (ID). Finally, we can see that it has also identified a pitch fall to signify word boundaries (WB).

The NNCGs are realised by feedforward neural networks with a single hidden layer and take as input four RU position ramps, outputting pitch and duration coefficients for each of the RUs, as shown in Fig. 1. The preferred RUs in SFC are the Inter Perceptual Centre Group (IPCG) units, that encompass the interval between two consecutive vocalic onsets [5]. This is advantageous for the analysis of the duration dynamics in speech [13].

The input ramps of the NNCGs indicate the absolute and relative position of the current syllable within the global scope of the linguistic function, as well the local scope if the function contains subscopes, e.g. a *left* and *right* subscope for a syntactic dependency between the verb (*left* scope) and object (right scope), where the function landmark would be placed on the RU that the verb ends with.

Pitch targets are output by the NNCGs at given time points within the vocalic nucleus of each RU, e.g. at 10%, 50% and 90% of its length [14]. A single duration coefficient is output per RU.

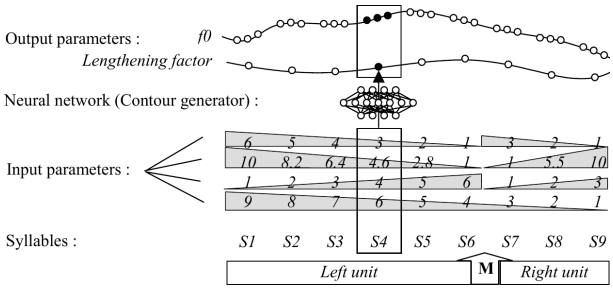


Figure 1: *Contour generator input parameters based on the scope of the function contour. Taken from [3].*

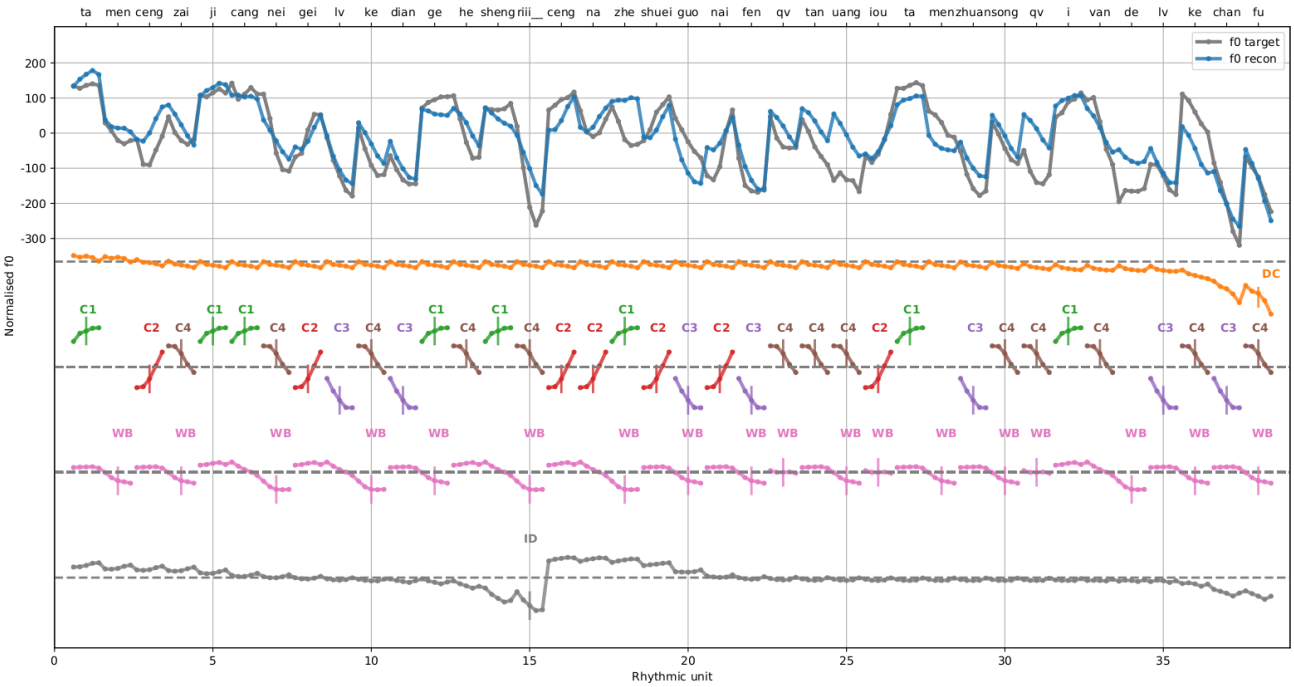


Figure 2: *Example PySFC decomposition of the intonation contour of the Chinese utterance: "Ta1 men ceng2 zai4 ji1 cang1 nei4 gei2 lv3 ke4 dian3 ge1 he4 sheng1 ri4, ceng2 na2 zhe1 shui2 guo3 nai2 fen3 qu4 tan4 wang4 you2 ta1 men zhuan3 song4 qu4 yi1 yuan4 de lv3 ke4 chan3 fu4." into constituent functional contours: declaration (DC), tones (C1-4), word boundaries (WB), and independent clauses (ID). Note that Tone 0 is not included in the decomposition.*

## 3. Objectives and State-of-the-art

The existence of the carry-over effect is well established in literature [15, 16, 17]. Due to articulatory inertia [18], in connected speech, a large portion of each tone consists of a movement away from the final state of the preceding tone and toward the underlying properties of the current tone [15].

It is our objective to validate these facts through the effects of scope on the modelling accuracy of the SFC. The SFC is able to separate the influences of the different elementary contours and thus does not require an especially tailored database for this purpose. We hypothesise that because of the established dominance of the carry-over effect over the anticipatory effect, the inclusion of the right syllable in the tone function scope will lead to significant improvements in intonation modelling, while the inclusion of the left one will conversely not.

## 4. Experiments

To improve the pitch modelling capabilities of SFC we have chosen to increase the number of targets in the vowel nucleus to 5, instead of 3. We evaluate our model using normalised $f_0$ expressed in semitones, which can be obtained using:

$$f_0[\text{semitones}] = 12 \log_2 \frac{f_0[\text{Hz}]}{f_R[\text{Hz}]} \tag{1}$$

where $f_R$ is the speaker specific reference pitch, i.e. the median. We have also decided to use syllables instead of IPCGs as RUs.

To assess the significance of the scope in the modelling of tonal contours in Chinese, we used the PySFC to change the possible scope of the functional contours of the tones. We used three settings:

- *narrow* scope – the scope of the tone function is limited to the syllable on which it is realised. This is the baseline setting,
- *left* scope – the scope includes the syllable preceding the syllable with the tone,
- *right* scope – the scope includes the syllable following the syllable with the tone,
- *double* scope – the scope includes both the syllable preceding and the one following the syllable with the tone.

Using these different settings we fit the SFC model to the data and analyse its performance using the weighted root mean square error (WRMSE), this :

$$WRMSE[\text{semitones}] = \sqrt{\frac{\sum_{n=0}^{N-1} w[n](f_0[n] - \hat{f}_0[n])^2}{\sum_{n=0}^{N-1} w[n]}}, \tag{2}$$

where $f_0$ is the original pitch contour, $\hat{f}_0$ is its reconstruction output by the SFC model, $w$ is the weighting function and $N$ is the number of samples in the utterance. Since the SFC works on predicting pitch targets within the vowel nucleus, we choose a binary $w$ that is 1 only within vowel regions, thus limiting the WMRSE calculation to $f_0$ in vowels.

The analysis was based on a database of read Chinese from a single speaker comprising 110 carrier utterances ranging from 6 to 38 syllables in length [9]. The database was annotated with the following functional contours: declaration, question, tones, word boundaries, interdependent and independent clauses. The number of occurrences of the four Chinese tones are given in Table. 1. We have chosen to leave Tone 0 out of our analysis as it does not encode a target intonation pattern.

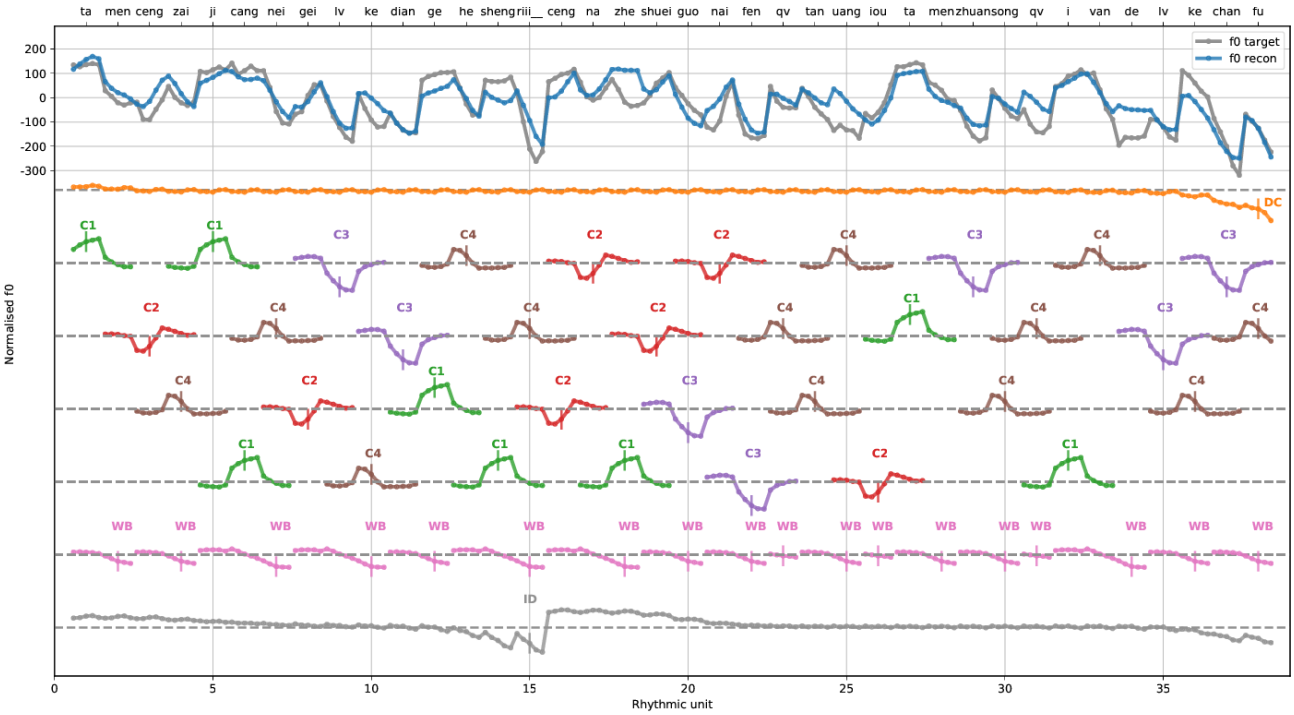


Figure 3: *PySFC decomposition of the same utterance with a* double *tone function scope.*

Table 1: *Distribution of the four tones in the database.*

| Tone | Count |
|---|---|
| Tone 1 | 793 |
| Tone 2 | 905 |
| Tone 3 | 563 |
| Tone 4 | 1146 |
| Tone 0 | 181 |

## 5. Results

Example PySFC decompositions of the same utterance shown in Fig. 2 using the *double* scope setting can be seen in Fig. 3. We can see that, compared to the *narrow* scope decomposition, the tone contours in the *double* scope decomposition are three times as long since they include both the previous and the following syllables in their scopes. The tone contour shapes for the carrier syllables in both cases are almost the same. Since, the *left* and *right* scope tone contours look like a truncated version of the *double* scope contours, they were not included for brevity.

In the *double* scope decomposition, we can also see that the tone contours have very low amplitudes for the neighbouring syllables, and realise the biggest part of their dynamics in the vowel nucleus of the carrier syllable. Also, the pitch movements associated with following syllable are somewhat greater than those of the preceding one. This can be seen better in the superposition of the tone contours for the *double* scope in Fig. 4.

The plot shows the dynamics of the contours for the carrier syllable, i.e. RU 0, as well as the preceding and following ones, i.e. RU -1 and RU 1. The pitch here is again expressed in 10s of quartertones. The contours' offsets have been adjusted so that all of them start from 0. To quantify this difference in pitch dynamics we calculated the mean squared pitch amplitude of the four tones within each of the three RUs in their scope. The results are given in Table 2. In contrast, the dynamics of the duration expansion coefficient for the tone functional contours are minimal as can be seen in Fig. 5.

Table 2: *Root mean squared pitch amplitude of the four tones within each of the three RUs in their scope.*

| Tone | RU -1 | RU 0 | RU 1 |
|---|---|---|---|
| Tone 1 | 5.57 | 96.96 | 18.37 |
| Tone 2 | 4.89 | 50.46 | 10.48 |
| Tone 3 | 5.15 | 109.33 | 32.22 |
| Tone 4 | 6.06 | 44.36 | 8.98 |
| Average | 5.42 | 75.28 | 17.51 |

Table 3: *WRMSE obtained for the different tone scope settings.*

| Scope | WRMSE [semitones] mean $\pm$ standard deviation |
|---|---|
| *narrow* | $2.92 \pm 0.38$ |
| *left* | $2.90 \pm 0.35$ |
| *right* | $2.85 \pm 0.38$ |
| *double* | $2.81 \pm 0.36$ |

Table 3 gives the mean and standard deviation of the WRMSE for the four different parameter choices for the tone scope. We can see that using a *narrow* scope gives the worse results, adding the *left* scope gives a slight improvement, while using the *right* scope and finally the *double* scope gives the best results. The distribution of the WRMSE across the files in the database is given using a boxplot in Fig. 6. We can see that the change in the scope parameter does not influence the spread of the distribution, as was evident in the standard deviation in Table 3, but only offsets the median.

To assess the significance of the change in WRMSE we conducted a repeated-measures analysis of variance (RM-ANOVA), since we were measuring the WRMSE on the same

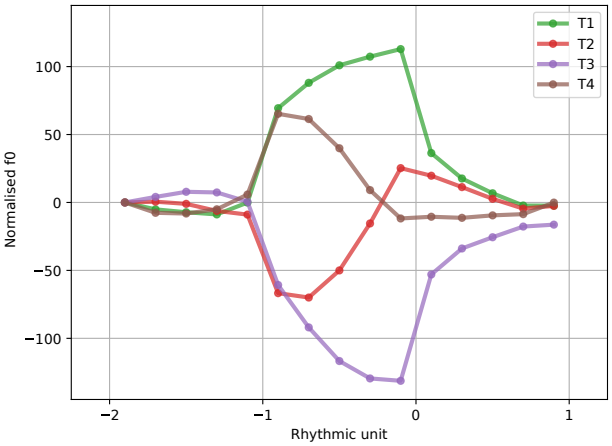


Figure 4: *A superposition plot of the $f_0$ contours generated by the 4 tone NNCGs using the* double *scope with both neighbouring syllables.*

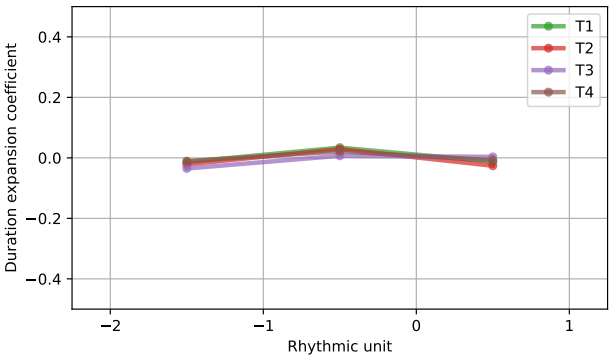


Figure 5: *A superposition plot of the duration coefficient contours generated by the 4 tone NNCGs using the* double *scope.*

Table 4: *Adjusted p-values obtained using a post hoc Tukey analysis for the different tone scope setting pairs.*

| Linear Hypothesis | $p$-value |
|---|---|
| *left - double* == 0 | <0.001 |
| *narrow - double* == 0 | <0.001 |
| *right - double* == 0 | 0.0251 |
| *narrow - left* == 0 | 0.5287 |
| *right - left* == 0 | <0.001 |
| *right - narrow* == 0 | <0.001 |

files in the different scope parameter settings. The analysis gave a $p$-value of $< .0001$, thus rejecting the null hypothesis. To find which parameter choice pairs did show statistical significance we carried out a Tukey post hoc analysis which gave the adjusted p values shown in Table 4.

## 6. Discussion

We can see that the only non-significant differences are *right-double* and *narrow-left*. This means that adding the left syllable in the *right* scope, i.e. when the right syllable was already added, and thus obtaining the *double* scope, does not bring statistically significant improvement to the WRMSE modelling results. Also, from the second non-significant pair we can conclude that adding the left syllable to the *narrow* scope does

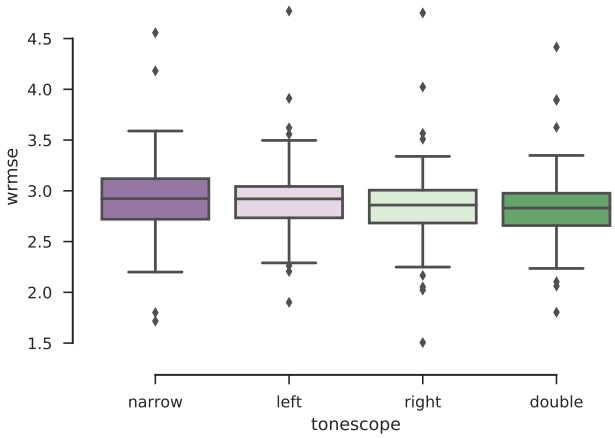


Figure 6: *Boxplot of the distribution of the WRMSE for the different scope parameter settings.*

not add statistically significant improvements of the modelling WRMSE. In other words, the addition of the left syllable in both cases is not statistically significant. This proves our initial hypothesis that the pitch movements associated with tones are not anticipatory, but rather carry-over to the next syllable.

## 7. Conclusions

We have demonstrated that the scope of the elementary intonation contours used to communicate the four tones of Chinese is not contained within the syllable that carries the tone, but in fact it also includes the following syllable. This carry-over effect might be due to the inertia of the laryngeal system that effectuates the pitch change. We have also shown that there is not a significant anticipatory component in the realisation of tones in our Chinese data.

The described methodology can be used to assess these effects in other tonal languages as well. Also, the PySFC can be used to explore the impact of neighbouring tone context on the tone shapes. A limitation of the proposed approach is that it can be applied only when the prosodic structure of the utterance matches its syntactical structure, on which the annotations and thus the input ramps of the NNCGs are based.

Future work will encompass the expansion of the prosody model to incorporate weight parameters that can be used to model contour prominence. This would allow the model to be used for example, to explore the impact of linguistic context on prominence.

## 8. Acknowledgements

This work has been conducted with the support of the Horizon 2020 Marie Skłodowska-Curie Actions Individual Fellowship Project, Call H2020-MSCA-IF-2016, under the project "ProsoDeep: Deep understanding and modelling of the hierarchical structure of Prosody".

## 9. References

[1] Yuen Ren Chao, *A Grammar of Spoken Chinese*, University of California Press, 1968.

[2] Moira Yip, *Tone*, Cambridge University Press, 2002.

[3] Gérard Bailly and Bleicke Holm, SFC: a trainable prosodic model," *Speech communication*, vol. 46, no. 3, pp. 348–364, 2005.

[4] B Holm and Gérard Bailly, Learning the hidden structure of intonation: implementing various functions of prosody," in *Speech Prosody 2002, International Conference*, 2002.

[5] Gérard Bailly and Bleike Holm, Learning the hidden structure of speech: from communicative functions to prosody," *Cadernos de Estudos Linguisticos*, vol. 43, pp. 37–54, 2002.

[6] Yann Morlec, Gérard Bailly, and Véronique Aubergé, Generating prosodic attitudes in French: data, model and evaluation," *Speech Communication*, vol. 33, no. 4, pp. 357–371, 2001.

[7] Yann Morlec, Albert Rilliard, Gérard Bailly, and Véronique Aubergé, Evaluating the adequacy of synthetic prosody in signalling syntactic boundaries: methodology and first results," in *Proceedings of the first International Conference on Language Resources and Evaluation. Granada, Spain*, 1998, pp. 647–650.

[8] Cécile Brichet and Véronique Aubergé, La prosodie de la focalisation en français: faits perceptifs et morphogénétiques," *Journées d'Etudes sur la Parole, Nancy-France*, pp. 33–36, 2004.

[9] Gao-Peng Chen, Gérard Bailly, Qing-Feng Liu, and Ren-Hua Wang, A superposed prosodic model for Chinese text-to-speech synthesis," in *Chinese Spoken Language Processing, 2004 International Symposium on*. IEEE, 2004, pp. 177–180.

[10] Branislav Gerazov and Gérard Bailly, PySFC – A System for Prosody Analysis based on the Superposition of Functional Contours Prosody Model," in *Speech Prosody*, 2018, submitted for review.

[11] Yann Morlec, *Génération multiparamétrique de la prosodie du français par apprentissage automatique*, Ph.D. thesis, 1997.

[12] Bleicke Holm and Gérard Bailly, Generating prosody by superposing multi-parametric overlapping contours.," in *INTERSPEECH*, 2000, pp. 203–206.

[13] W Nick Campbell, Syllable-based segmental duration," *Talking machines: Theories, models, and designs*, pp. 211–224, 1992.

[14] Stéphanie de Tournemire, Identification and automatic generation of prosodic contours for a text-to-speech synthesis system in French," in *Fifth European Conference on Speech Communication and Technology*, 1997, pp. 191–194.

[15] Yi Xu, Contextual tonal variations in Mandarin," *Journal of phonetics*, vol. 25, no. 1, pp. 61–83, 1997.

[16] Yi Xu and Q Emily Wang, Pitch targets and their realization: Evidence from Mandarin Chinese," *Speech communication*, vol. 33, no. 4, pp. 319–337, 2001.

[17] Yiya Chen and Carlos Gussenhoven, Emphasis and tonal implementation in Standard Chinese," *Journal of Phonetics*, vol. 36, no. 4, pp. 724–746, 2008.

[18] Yi Xu and Xuejing Sun, Maximum speed of pitch change and how it may relate to speech," *The Journal of the Acoustical Society of America*, vol. 111, no. 5, pp. 1399–1413, 2002.